WME: Sense, Polarity and Affinity based Concept Resource for Medical Events

[1]Anupam Mondal **[1]Dipankar Das** **[2]Erik Cambria** **[1]Sivaji Bandyopadhyay**
[1]Department of CSE
Jadavpur University, India
[2]School of Computer Engineering
Nanyang Technological University, Singapore
[1]link.anupam@gmail.com,[1]ddas@cse,jdvu.ac.in,
[2]cambria@ntu.edu.sg,[1]sbandyopadhyay@cse.jdvu.ac.in

Abstract

In order to overcome the lack of medical corpora, we have developed a WordNet for Medical Events (WME) for identifying medical terms and their sense related information using a seed list. The initial WME resource contains 1654 medical terms or concepts. In the present research, we have reported the enhancement of WME with 6415 number of medical concepts along with their conceptual features viz. Parts-of-Speech (POS), gloss, semantics, polarity, sense and affinity. Several polarity lexicons viz. SentiWordNet, SenticNet, Bing Liu's subjectivity list and Taboda's adjective list were introduced with WordNet synonyms and hyponyms for expansion. The semantics feature guided us to build a semantic co-reference relation based network between the related medical concepts. These features help to prepare a medical concept network for better sense relation based visualization. Finally, we evaluated with respect to Adaptive Lesk Algorithm and conducted an agreement analysis for validating the expanded WME resource.

1 Introduction

In the domain of clinical text processing, sense-based information extraction is considered as a challenging task due to the unstructured nature of the corpus. The difficulty in preparing structured corpora from unstructured corpora in clinical domain is apparent due to the lack of involvement from the domain experts (e.g. medical practitioners) (Smith and Fellbaum, 2004). Several lexicons or systems were developed and used by researchers to overcome the above-mentioned difficulty in the conventional Natural Language Processing (NLP) domain (Miller, 1995; Fellbaum, 1998). In contrast, the researchers in the medical domain have introduced some resources e.g., Medical WordNet to overcome such a problem (Burgun and Bodenreider, 2001; Bodenreider et al., 2003). The WME resource was developed along with sense-based medical information for the use of experts like medical practitioners and non-experts like patients (Mondal et. al., 2015).

In the present attempt, we have expanded the WME resource with new features like semantics and affinity. Semantics feature helps to extract the relative sense-based words of the medical concepts from different knowledge bases and assign the medical words to their appropriate categories (e.g. treatment, disease, etc.). Affinity feature helps to develop a medical Concept Network (ConceptNet) for visualizing the concept relations (Cambria et al., 2010). Starting with an initial seed list of medical concepts, the synonyms and hyponyms of the concepts in the WordNet along with several polarity lexicons were extracted to enrich the present version of WME. The polarity lexicons in SentiWordNet[1], SenticNet[2], Bing Liu's subjectivity list[3] and Taboda's adjective list[4] were applied on the extracted synonyms and hyponyms so as to identify the proper sense of the retrieved medical concepts.

Section 2 provides an overview of our research. Section 3 illustrates the related work associated to preparation of lexical resources in the clinical domain. Section 4 discusses WME expansion techniques along with related tabulations needed for WME building. Section 5 provides important features selection and identification approaches of WME. Section 6 describes the process taken to evaluate the expanded WME resource along with agreement studies. Finally Section 7 gives a conclusion to our research and mentions the future plans of our study.

[1]http://sentiwordnet.isti.cnr.it/
[2]http://sentic.net/
[3]http://www.cs.uic.edu/~liub/
[4]https://www.sfu.ca/~mtaboada/research/pubs.html

2 Overview

Sentiment-oriented information extraction system or lexicon preparation is treated as a contributory research in NLP due to lack of involvement from domain expert (e.g. medical practitioners). Specifically in the domain of Biomedical Natural Language Processing, the extraction of medical concepts and their related sense, polarity and semantic feature is difficult due to the unstructured nature of the medical or clinical corpora. In this research, we aim to overcome these above-mentioned challenges by expanding the WordNet of Medical Event (WME) through including knowledge-based features. Semantic and affinity features help us to prepare an affinity relation-based network, known as semantic and concept networks, for the WME. To enhance the WME resource, we have primarily concerned with the following subsections:

1. Feature Selection for WME expansion (Section 5.1)

2. Evaluation (Section 6.1)

3. Agreement Analysis (Section 6.2)

3 Related Work

In the context of Biomedical corpora, the medical concepts (events) and their related information extraction can help to develop an annotation system, which is essential to build structured medical corpora (UzZaman and Allen, 2010; Hogenboom et al., 2011). The polarity, sense and concept related features are crucial for preparing the structured corpus in this domain.

Several taxonomies were designed by researchers to allow non-experts (e.g. patients) to better understand medical concepts and their related information (Tse, 2003; Zeng et al., 2003). In this concern, Patel et al. (2002) developed a medical information system by compiling a list of medical vocabulary and provide the context of the medical words as understood by experts and non-experts. Fellbaum and Smith developed Medical WordNet (MEN) along with two sub networks namely, Medical FactNet (MFN) and Medical BeliefNet (MBN), which serves as a source of consumer health information that provides medical information explanation to patients (Smith and Rosse, 2004). MEN were developed under the formal architecture of the Princeton WordNet (Fellbaum, 1998). MFN helps the non-expert group to extract and comprehend generic medical information, whereas the MBN identifies the fraction of the beliefs about medical phenomenon (Smith and Rosse, 2004). Their primary motivation was to develop a medical information retrieval system with visualization effects.

The extraction of medical terms from the clinical corpus is an ambiguous task (Pustejovsky, 1995). Therefore, a group of researchers have introduced sense selection and pruning strategies to expand the ontology of the medical domain (Toumouh et al., 2006). WordNet of Medical Event (WME) was introduced as a lexical resource to identify medical events and their related features viz. POS, gloss, polarity and sense from the corpora (Mondal et. al., 2015). The POS of the medical concepts signifies the lexical categories of the medical events, whereas their gloss, polarity and sense features help to provide the semantics and knowledge-based information relating to the medical events.

4 WME1.0 Building

Sense-based keyword extraction is essential for context sense identification (e.g. In the sentence “A supplementary component that improves capability”, the keywords “*improves*” and “*capability*” keywords denote the positive sense of the sentence). It is a tedious job in Biomedical Natural Language Processing domain, and it is because knowledge-based meaning identification along with the POS, synonyms, hyponyms and definition of the words has to be extracted from the conventional WordNet. But such an approach is not adequate to provide appropriate knowledge and sense-based information needed for the medical concepts (terms). To identify the syntactic and semantic features of the medical concepts, we have developed WME1.0 that provides the POS, gloss and sense of the medical terms. The seed list of WME1.0 resource was prepared from the trial and training datasets of the SemEval-2015 Task-6[5]. In addition to the conventional WordNet, we have also used English medical dictionary to develop the initial WME resource. SemEval 2015 Task-6 datasets have extracted 2479 medical events along with their attributes such as *type*,

[5]http://alt.qcri.org/semeval2015/task6/

span-context, sense (*positive/negative*) from the provided datasets (e.g., *<tumor>, <event>, <An abnormal new mass of tissue that serves no purpose.>, <negative>*). The POS, synonyms and definition of the seed list were then added from the WordNet (e.g., *<Abdomen>, <Noun>, <1. abdomen 2. abdominal cavity>, <1. "The region of the body is vertebrate between the thorax and the pelvis." 2."The cavity containing the major viscera; in mammals it is separated from the thorax by the diaphragm.>*). Meanwhile, an English Medical Dictionary[6] identifies the POS and word related gloss of these medical concepts. From the above-mentioned dictionary, we have to perform manual editing for preprocessing on 11750 medical words in English along with their POS and gloss (e.g., *<Adenoma>, <Noun>, <A benign tumor of a gland>*).

In order to identify the proper sense-based gloss of the seed list, we have used SenticNet, SentiWordNet, Bing Liu subjective list and Taboda's adjective list sentiment lexicons. After extracting various sense-based glosses from different resources, we chose the knowledge-based gloss that is most appropriate to the medical concepts by introducing (Mondal et al., 2015) sequential and combined Word Sense Disambiguation (WSD) (Basili et al., 1997).

5 WME2.0 Building

The inclusion of semantic and knowledge-based features is crucial for preparing an expanded version of the WME. The semantic, polarity, sense and affinity features have been introduced to identify and extract medical concepts (events) from the clinical corpora. In the following subsections, we have discussed in details the steps taken for features selection and their related statistical observations.

5.1 Feature Selection for Expansion

In order to better understand the concepts in the new version of WME, sense and knowledge feature selection is more important than sense-based matching. This is because sense and knowledge (polarity and semantics) features help to visualize the relationship between these concepts through affinity scores based on co-reference relations of the medical concepts (Cambria et al., 2015). The following features are taken into account to design the new version of WME resource.

Gloss: To identify the syntactical knowledge-based information of the medical concepts, the descriptive gloss is essential for evaluating the meaning of the concept. Specifically, if the gloss of a concept has been collected from different resources, it is challenging to identify the proper gloss appropriate for the medical context due to various competing senses. For the proper identification of sense-base gloss in WME, we have proposed two Word Sense Disambiguation (WSD) approaches, namely Sequential and Combined. These approaches help to identify the sense-based glosses of the seed list of WME2.0 resource that are appropriate in the medical context (Mondal et. al., 2015).

Polarity and Sense: In the medical field, sentiment or opinion extraction is a burgeoning field due to the lack of available sentiment resources for such a domain. We attempted to overcome this problem by introducing polarity and its related sense features in our WME resource. We have considered several polarity lexicons viz. SentiWordNet, SenticNet, Bing Liu's subjective list and Taboda's adjective list to extract the polarity and sense features of the medical concepts. Figure 1 shows the procedures taken by WME2.0 to identify the polarity and sense features of a particular concept (e.g., <Concept: mismanage>, <Polarity: -0.625>, <Sense: Negative>).

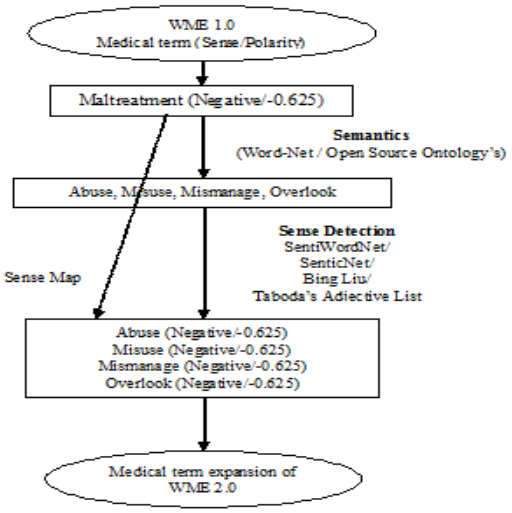

Figure 1. Diagram of sense-based technique in WME 2.0

[6]http://alexabe.pbworks.com/f/Dictionary+of+Medical+Terms+4th+Ed.-+(Malestrom).pdf

Semantic: Medical concept and the identification of their sense-related words are essential to prepare a medical semantic network, which can help to explain how those concepts are related to each other. To annotate the medical concepts with their relevant semantic features, we have utilized the synonyms, hyponyms provided by WordNet along with different knowledge-based resources like SenticNet for the expansion of WME2.0 (e.g., <Concept: maltreatment>, <Semantics: *abuse*, *misuse*, *mismanage*, *overlook*>). The additional semantic features provided by these resources help to analyze the meaning of the concepts and their relations with the glosses.

Affinity: Affinity indicates the natural link between similar concepts or words. The degree or affinity score of the medical concept help to develop concept clusters in WME. These concept clusters then help to build a semantic and concept networks for the better understanding and visualization of the concepts and their relations. For example, the medical concept "brain" has affinity score of 0.0290 with "alive", a score of 0.1540 with "clog", and 0.0560 with "fall in love". These scores indicate the degree of relation between these concepts. The affinity score of these concepts is calculated by using a probabilistic approach. Equation (1) shows the computing process of $Affinity_{(c)}$.

$$Affinity_{(c)} = MT_{1(c)} \cap MT_{2(c)} \quad (1)$$

where $MT_{1(c)}$ and $MT_{2(c)}$ denote two different medical concepts.

From the extracted $Affinity_{(c)}$, the Affinity score ($Affinity\text{-}Score_{(c)}$) between two concepts is then calculated with Equation (2).

$$Affinity\text{-}Score_{(c)} = Affinity_{(c)} / \sum MT_{i\,(c)} \quad (2)$$

where $i=2$ indicates the two semantic sets, namely $MT_{1(c)}$ and $MT_{2(c)}$. $Affinity_{(c)}$ indicates the number of semantics in common with these medical concepts.

The $Affinity\text{-}Score_{(c)}$ shows the co-reference relation between these medical concepts, which can range from 0 to 1. Figure 2 shows the partial representation of the semantic network that illustrates the relations between the medical concepts based from their affinity scores.

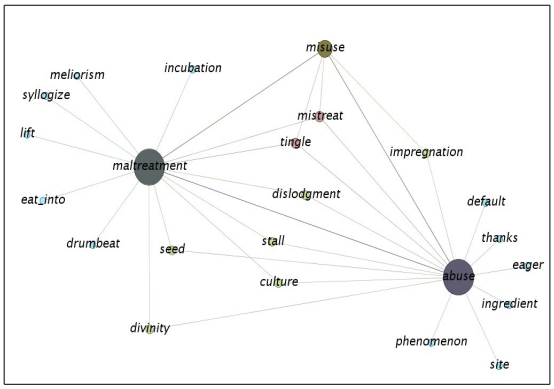

Figure 2. Relations of the concepts based on their affinity scores in the partial visualization of the semantic network

5.2 Tabulations

Table 1 shows the number of medical concepts, POS and sense distributions in our initial and expanded versions of WME, known as WME1.0 and WME2.0 respectively throughout this paper.

Different Basic Operation		**WME1.0**	**WME2.0**
No. of Medical Concepts		1654	6415
POS Distribution	Noun	1019	4219
	Verb	488	2026
	Adjective	124	111
Sense Distribution	Positive	1338	2800
	Negative	316	3615

Table 1. Comparative Tabulations

The table above indicates that it is difficult to expand the WME resource with the help of word level lexical analysis (POS distribution) due to the unavailability of appropriate medical lexicons. We have enlisted the help of the sentiment-based approaches to overcome such a difficulty through utilizing SentiWordNet, SenticNet, Bing Liu and Taboada's adjective list sentiment lexicons. Table 2 provides a detailed breakdown of the expanded medical concepts that has been elaborated by the above-mentioned sentiment lexicons along with their combined polarity lexicon. The combined polarity lexicon represents medical concepts that commonly occur in all of the above-mentioned lexicons.

		SW	SN	BL	TA	CM
O	**S**	2938	210	1250	2509	6698
	H	4125	1136	5301	9901	19328
U	**S**	1151	196	615	1017	1592
	H	1623	698	2761	4833	6584

SW → SentiWordNet, **SN** → SenticNet, **BL** → Bing Liu's subjectivity list, **CM** → Combined Medical List, **TA** → Taboda's Adjective List
O → Original terms **U** → Unique terms
S → Synonyms **H** → Hyponyms

Table 2. Tabulations based on Senses of different Polarity lexicons

Table 2 shows the polarity given by the Taboada's adjective list, Bing Liu's subjective list and SentiWordNet polarity lexicons for the expansion of the WME resource, whereas SenticNet (Cambria et al., 2014; Cambria E., 2016) introduces semantic feature of the medical terms in WME2.0.

6 Discussion

6.1 Evaluation

In contrast to WME1.0, we have performed the preliminary expansion of WME2.0 with the help of sense feature. The gloss sense of the medical concepts of WME2.0 was compared with the sense extracted from SentiWordNet lexicon. While developing our clinical corpus, we found that SentiWordNet is limited by the lack of medical concepts or words. We have observed that SentiWordNet only covers nearly 40% of the medical concepts present in WME2.0 resource. From the data extracted from various knowledge databases, we have evaluated the extracted glosses of the concepts and determined their proper senses in the medical context by using Lesk WSD algorithm over WME2.0. The simplified versions of Lesk algorithm mainly compare the extracted glosses with dictionary definition and generate the sense-based output of the medical concept. The simplified versions of Lesk algorithm, however, are not effective because of the insufficient number of medical concepts in their dictionary. To resolve this problem, we have enlisted the help of Adaptive Lesk algorithm to validate the sense-based descriptions. The Adaptive Lesk algorithm not only compares the extracted glosses with dictionary definitions, it also looks at synonymous set definitions in the WordNet. After evaluating the WME2.0 with Adaptive Lesk algorithm, we have calculated the F-Measure score for the medical concepts. Equation (3) shows how Recall (R) and Precision (P) help to calculate the F-Measure score of WME2.0.

$$F\text{-}Measure = 2 * [(R * P) / (R + P)] \quad (3)$$

For the identification of the sense-based medical concepts gloss, we observed the F-Measure score for WME2.0 and Adaptive Lesk approach are 0.71 and 0.38 respectively. In this process, the Precision and Recall scores for these approaches are 0.82 and 0.57 for WME2.0, and 0.62 and 0.29 for Adaptive Lesk approach respectively. The evaluation shows that WME2.0 provides more accurate sense-based gloss information in comparison to Adaptive Lesk algorithm for the medical concepts.

6.2 Agreement Analysis

We also have conducted a manual evaluation on top of statistical approaches to validate WME2.0. Manual annotators-based agreement analysis has to be conducted due to the unavailability of medical sense-based lexicons. Cohen's kappa (Cohen, 1960) based statistical approach has been used to calculate the accuracy of the agreement analysis, as shown in Equation (4). The Cohen's Kappa (k) score is measured by the Proportionate (Pr(a)) and Random (Pr(e)) agreement scores.

$$k = [Pr(a) - Pr(e)] / [1 - Pr(e)] \quad (4)$$

Table 3 shows the number of agreed (Y) and non-agreed (N) medical concepts and their related features by the two manual annotators (A and B). The number of agreed and non-agreed medical concepts by the annotators was then used to calculate the agreement score, namely Kappa value (k). The evaluated Kappa score of 0.73 provides a satisfactory output for the WME2.0 resource.

No. of Medical Terms 6415		**B**	
		Y	**N**
A	**Y**	6094	51
	N	77	193

Table 3. Agreement study of WME 2.0

7 Conclusion and Future Work

In the present research, we have expanded the WME resource by including syntactical and

semantic features to the medical concepts. In this concern, we have expanded the sense of medical concepts in our seed list through utilizing several sentiment lexicons along with their synonyms and hyponyms in the conventional WordNet. The new WME contains 6415 medical concepts along with their POS, gloss, semantics, polarity, sense and affinity features. Affinity and semantic features helps us to build a medical semantic network with co-reference relation between these medical concepts for the experts and non-experts group of people.
In future, we will attempt to enrich the WME2.0 resource by including more medical concepts along with their concept-based and knowledge-based features so as to improve the quality as well as coverage of the resource.

References

Mondal. A., Chaturvedi. I., Bajpai. R., Das. D., Bandyopadhyay. S. 2015. *Lexical Resource for Medical Events: A Polarity Based Approach.* IEEE 15th International Conference on Data Mining Workshops, Atlantic City.

Basili. R., DellaRocca. M. and Pazienza. M. T. 1997. *Contextual word sense tuning and disambiguation.* Applied Artificial Intelligence. pp. 235-262.

Bodenreider. O., Burgun. A. and Mitchell. J. A. 2003. *Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: a feasibility study.* Studies in Health Technology and Informatics. pp. 379-384.

Burgun. A. and Bodenreider. O. 2001. *Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System.* In: NAACL Workshop on WordNet and Other Lexical Resources, pp. 77-82.

Cambria. E., Hussain. A., Havasi. C., Eckl. C. 2010. *SenticSpace: Visualizing opinions and sentiments in a multi-dimensional vector space.* In: LNAI, vol. 6279, pp. 385–393.

Cambria. E., Fu. J., Bisio. F. and Poria. S. 2015. *AffectiveSpace 2: Enabling affective intuition for concept-level sentiment analysis*. In: AAAI, pp. 508-514, Austin.

Cambria. E., Olsher. D., and Rajagopal. D. 2014. *SenticNet 3: A common and common-sense knowledge base for cognition-drive sentiment analysis*. In: AAAI, pp. 1515-1521, Quebec.

Cambria. E. 2016. *Affective computing and sentiment analysis.* In: IEEE Intelligent Systems 31(2).

Cohen. J. 1960. *A coefficient of agreement for nominal scales.* In: Educational and Psychological Measurement, 20 (1), pp. 37-46.

Fellbaum. C. 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge.

Hogenboom. F., Frasincar. F., Kaymak. U. and de-Jong. F. 2011. *An overview of event extraction from text.* In: Derive Workshop, Bonn.

Miller. G. A. 1995. *WordNet: a lexical database for English.* Comm ACM. pp. 39-41.

Patel. V. L., Arocha. J. F. and Kushniruk. 2002. *A Patients' and physicians' understanding of health and biomedical concepts: relationship to the design of EMR systems.* Journal of Biomedical Informatics: 35(1). pp. 8-16.

Pustejovsky. J. 1995. *The generative lexicon.* MIT Press, Cambridge.

Smith. B. and Fellbaum. C. 2004. *Medical WordNet: A New Methodology for the Construction and Validation of Information Resources for Consumer Health.* In: Coling, Geneva, pp. 31-38.

Smith. B. and Rosse. C. 2004. *The role of foundational relations in the alignment of biomedical ontologies.* In: Medinfo, San Francisco.

Toumouh. A., Lehireche. A., Widdows. D. and Malki. M. 2006. *Adapting WordNet to the Medical Domain using Lexicosyntactic Patterns in the Ohsumed Corpus.* IEEE/ACS International Conference on Computer Systems and Applications (AICCSA).

Tse. A. Y. 2003. *Identifying and characterizing a consumer medical vocabulary.* Doctoral dissertation, College of Information Studies, University of Maryland, College Park.

UzZaman. N. and Allen. J. F. 2010. *Extracting Events and Temporal Expressions from Text.* Proceedings of IEEE International Conference on Semantic Computing.

Zeng. Q., Kogan. S., Ash. N., Greenes. R. A. and Boxwala. A. A. 2003. *Characteristics of consumer terminology for health information retrieval: A formal study of use of a health information service.* Methods of Information in Medicine.